CUNY SPS DATA 621 - CTG5 - HW4

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 24th, 2019

Contents

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET_FLAG	car crash = 1, no car crash = 0	response
TARGET_AMT	car crash cost = >0, no car crash = 0	response
AGE	driver's age - very young/old tend to be risky	numerical predictor
BLUEBOOK	$ value of vehicle	numerical predictor
CAR_AGE	age of vehicle	numerical predictor
CAR_TYPE	type of car (6types)	categorical predictor
CAR_USE	usage of car (commercial/private)	categorical predictor
CLM_FREQ	number of claims past 5 years	numerical predictor
EDUCATION	max education level (5types)	categorical predictor
HOMEKIDS	number of children at home	numerical predictor
HOME_VAL	$ value of home - home owners tend to drive more responsibly	numerical predictor
INCOME	$ income - rich people tend to get into fewer crashes	numerical predictor
JOB	job category (8types, 1missing)- white collar jobs tend to be safer	categorical predictor
KIDSDRIV	number of driving children - teenagers likely get into crashes	numerical predictor
MSTATUS	maritial status - married people drive more safely	catogerical predictor
MVR_PTS	number of traffic tickets	numerical predictor
OLDCLAIM	$ total claims in the past 5 years	numerical predictor
PARENT1	single parent	categorical predictor
RED_CAR	a red car	categorical predictor
REVOKED	license revoked (past 7 years) - more risky driver	categorical predictor
SEX	gender - woman may have less crashes than man	categorical predictor
TIF	time in force - number of years being customer	numerical predictor
TRAVTIME	distance to work	numerical predictor
URBANCITY	urban/rural	categorical predictor
YOJ	years on job - the longer they stay more safe	numerical predictor

1 DATA EXPLORATION

In this assignment, we explore, analyze and model a dataset containing 8,161 rows and 25 columns. Of all 25 features, 14 are discrete and 11 are continuous. There are total 970 missing values out of 204,025 observations.

We will build a logistic and multiple linear regression that will determine the followings: • Predict the probability that a person will crash their car • Amount of money it will cost if the person does crash their car

We will be able to develop insurance rates based on a number of predictors such as income, age, distance to work, and how long they have been customers, etc.

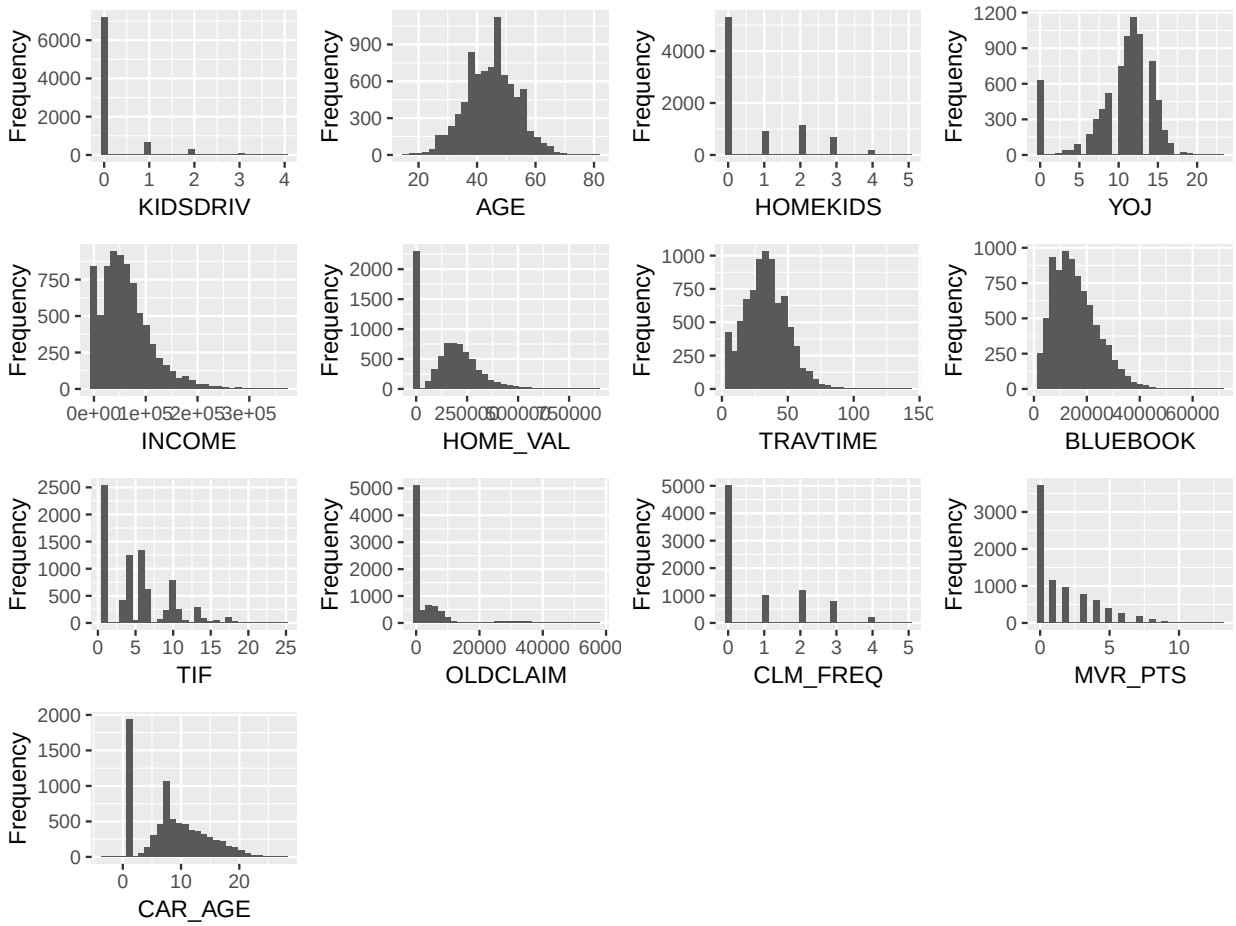
In the training dataset, there are 23 predictors and 2 response variables – one is binary value that indicates whether claim was made and the other is numerical value indicating the cost of claim.

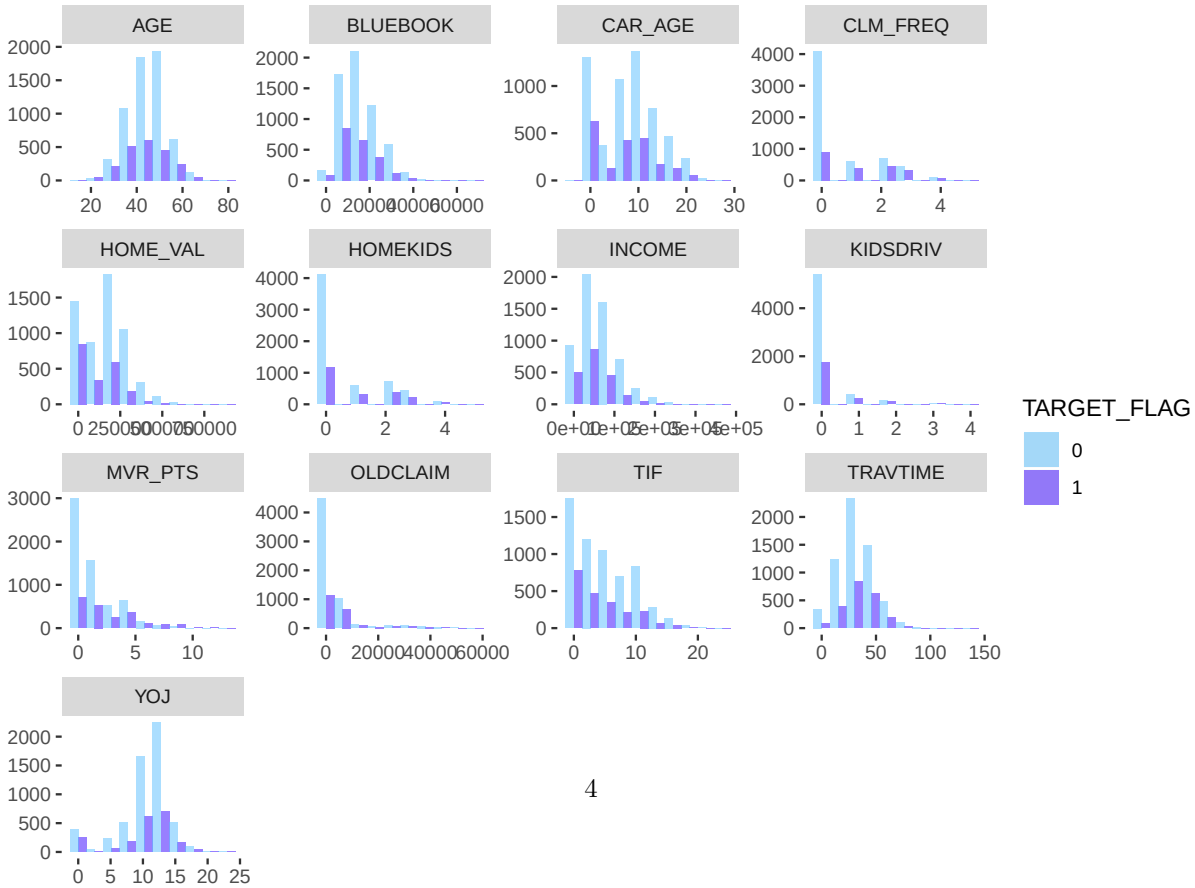
The response variable shows appropriate distribution in the training data. We confirm that for the number of target flags are 0 equals the target amount 0.

Table 2: Summary statistics

	n	mean	sd	median	min	max	skew	kurtosis
KIDSDRIV	8161	1.710575e-01	5.115341e-01	0	0	4	3.3518374	11.7801916
AGE	8155	4.479031e+01	8.627589e+00	45	16	81	-0.0289889	-0.0617020
HOMEKIDS	8161	7.212351e-01	1.116323e+00	0	0	5	1.3411271	0.6489915
YOJ	7707	1.049929e+01	4.092474e+00	11	0	23	-1.2029676	1.1773410
INCOME	7716	6.189809e+04	4.757268e+04	54028	0	367030	1.1863166	2.1290163
HOME_VAL	7697	1.548673e+05	1.291238e+05	161160	0	885282	0.4885950	-0.0160838
TRAVTIME	8161	3.348572e+01	1.590833e+01	33	5	142	0.4468174	0.6643331
BLUEBOOK	8161	1.570990e+04	8.419734e+03	14440	1500	69740	0.7942141	0.7913559
TIF	8161	5.351305e+00	4.146635e+00	4	1	25	0.8908120	0.4224940
OLDCLAIM	8161	4.037076e+03	8.777139e+03	0	0	57037	3.1190400	9.8606583
CLM_FREQ	8161	7.985541e-01	1.158453e+00	0	0	5	1.2087985	0.2842890
MVR_PTS	8161	1.695503e+00	2.147112e+00	1	0	13	1.3478403	1.3754900
CAR_AGE	7651	8.328323e+00	5.700742e+00	8	-3	28	0.2819531	-0.7489756

1.1 Summary Statistics





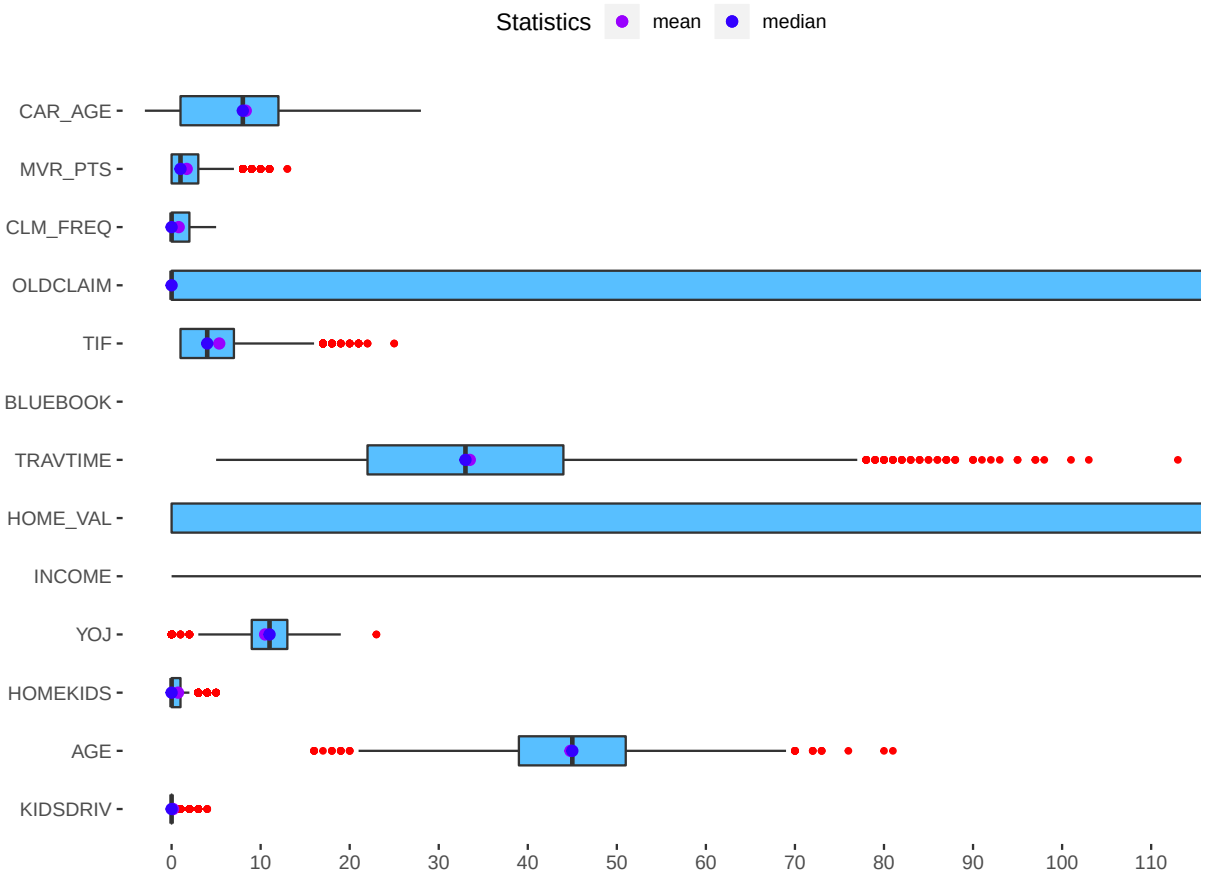


Figure 1: Outliers Boxplot

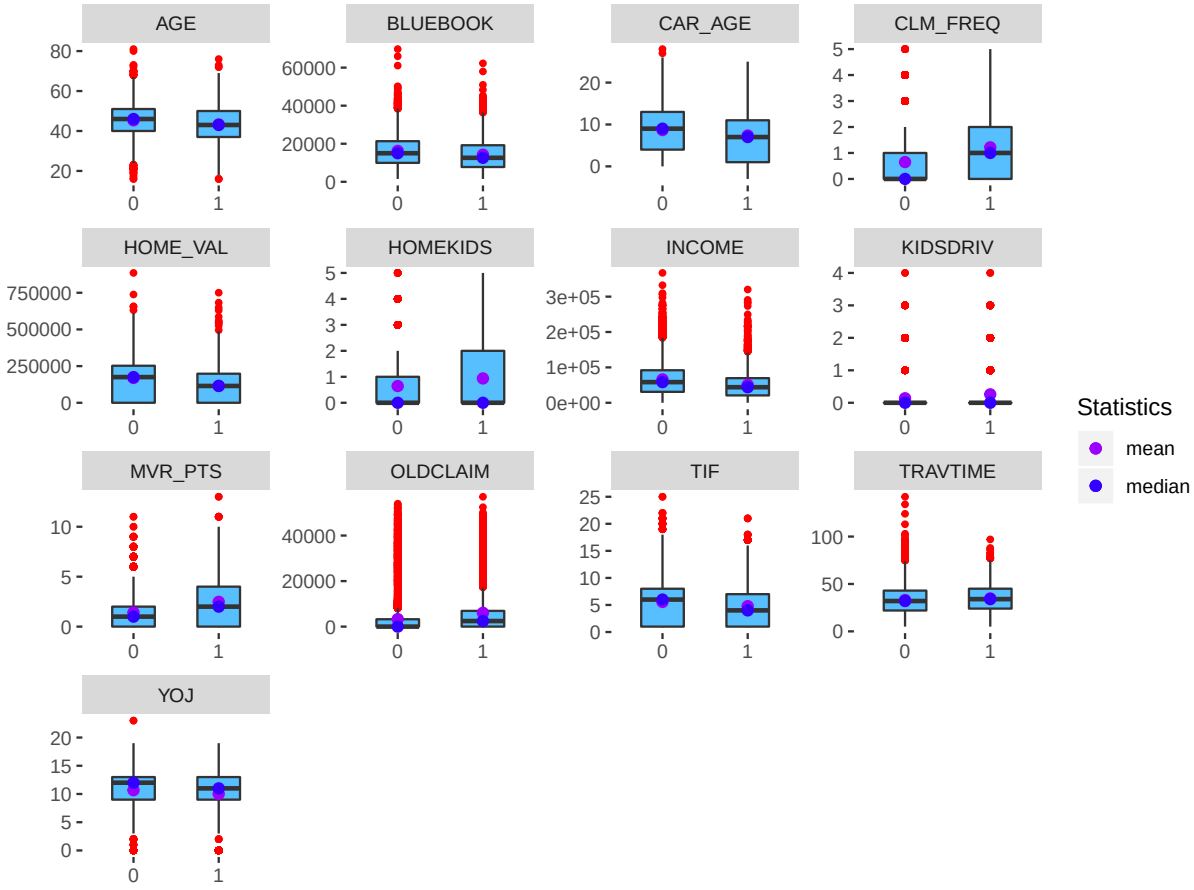
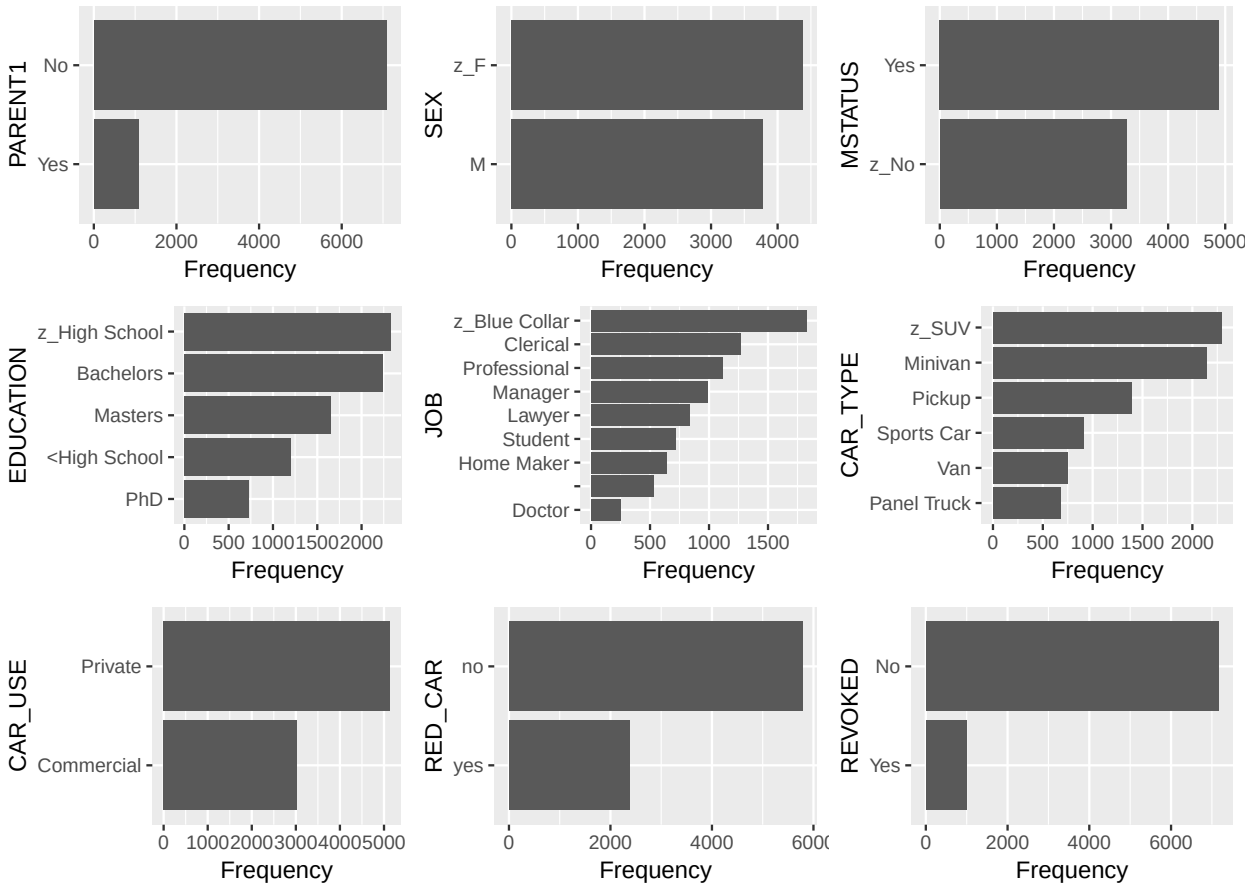
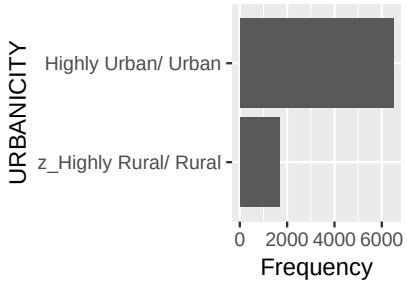


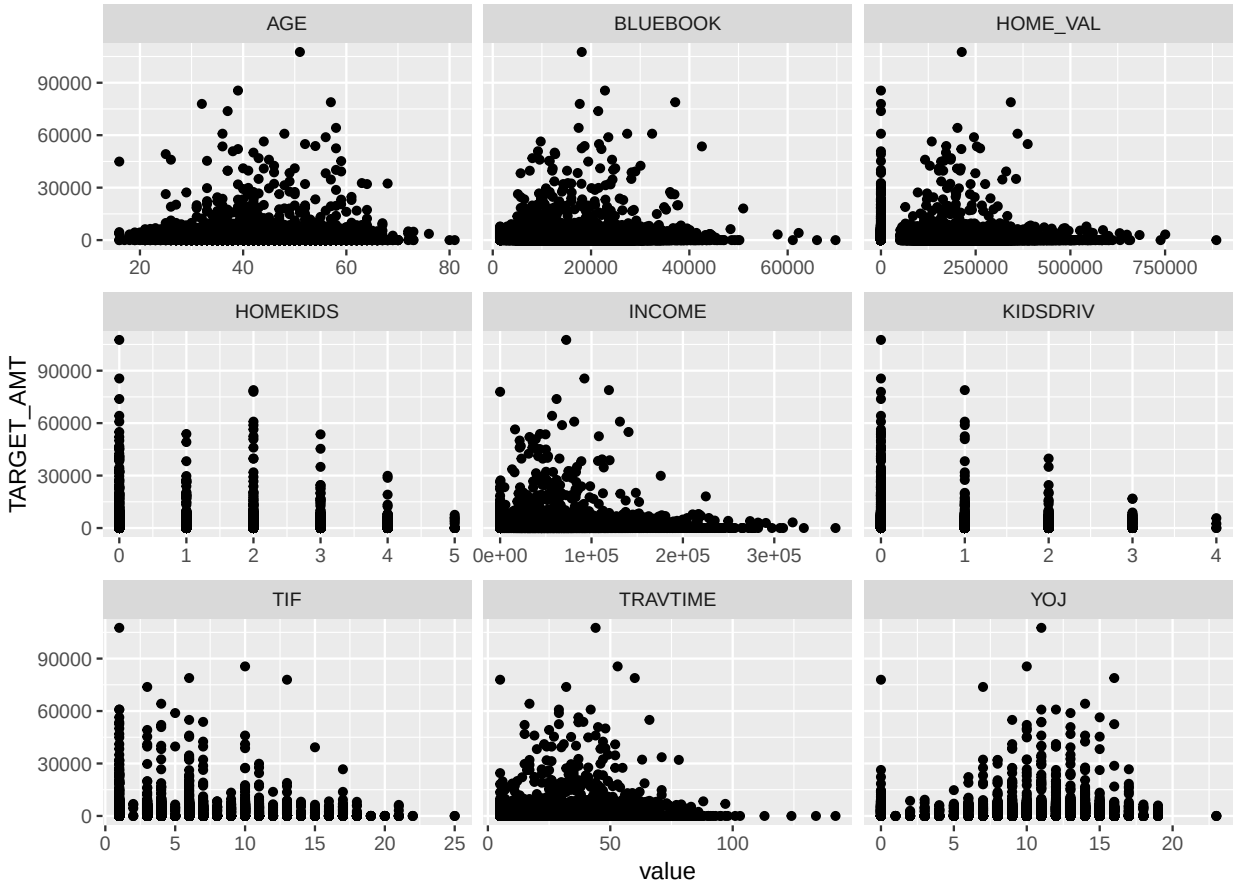
Figure 2: Linear relationship between each numeric predictors and the target


PARENT1
No
Yes
0 2000 4000 6000
Frequency
SEX
z_F
M
0 1000 2000 3000 4000
Frequency
MSTATUS
Yes
z_No
0 1000 2000 3000 4000 5000
Frequency
EDUCATION
z_High School
Bachelors
Masters
<High School
PhD
0 500 1000 1500 2000
Frequency
JOB
z_Blue Collar
Clerical
Professional
Manager
Lawyer
Student
Home Maker
-
Doctor
0 500 1000 1500
Frequency
CAR_TYPE
z_SUV
Minivan
Pickup
Sports Car
Van
Panel Truck
0 500 1000 1500 2000
Frequency
CAR_USE
Private
Commercial
0 1000 2000 3000 4000 5000
Frequency
RED_CAR
no
yes
0 2000 4000 6000
Frequency
REVOKED
No
Yes
0 2000 4000 6000
Frequency

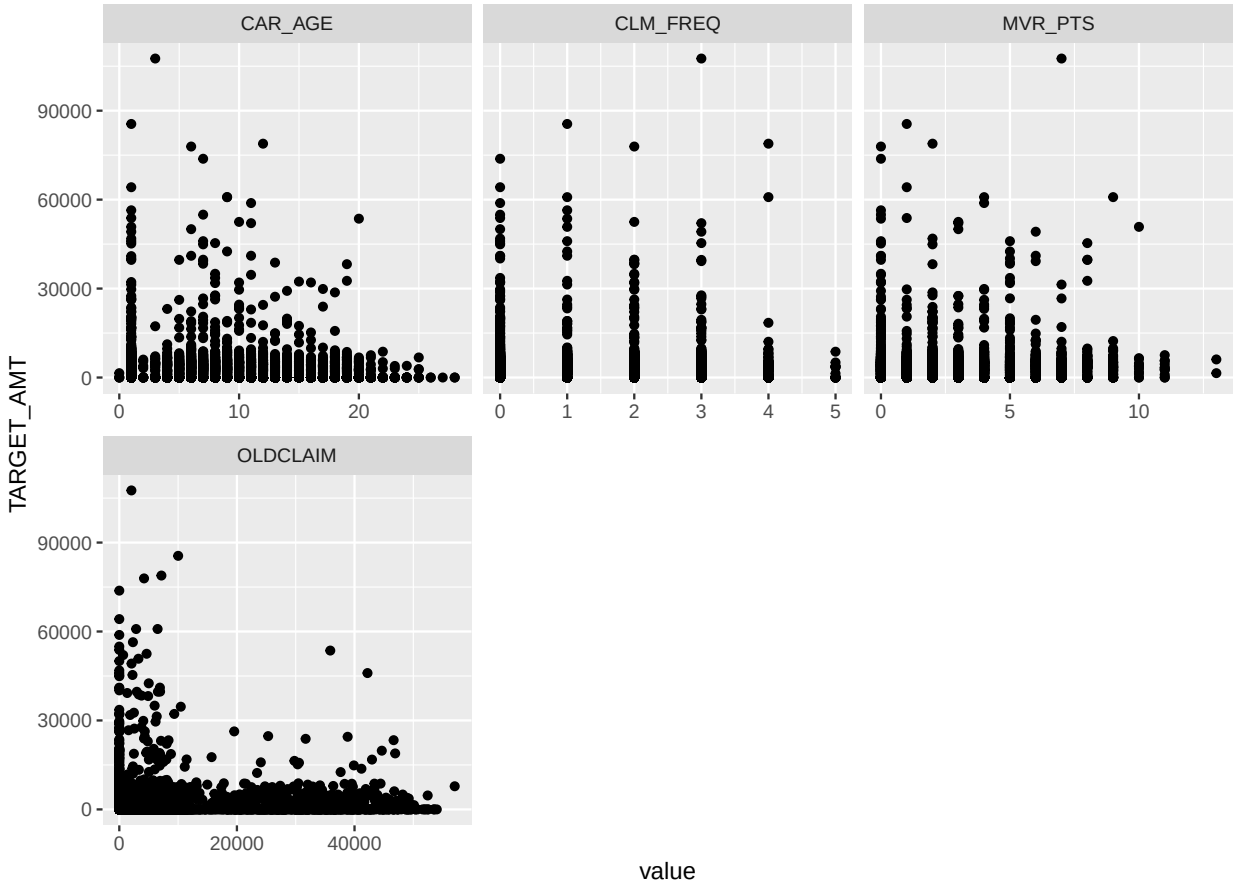
Page 1


URBANICITY
Highly Urban/ Urban
z_Highly Rural/ Rural
0
2000
4000
6000
Frequency

Page 2


AGE
BLUEBOOK
HOME_VAL
HOMEKIDS
INCOME
KIDSDRIV
TIF
TRAVTIME
YOJ
TARGET_AMT
value
90000
60000
30000
0
20
40
60
80
0
20000
40000
60000
250000
500000
750000
1
2
3
4
5
0e+00
1e+05
2e+05
3e+05
10
15
20
25
50
100

Page 1


CAR_AGE
CLM_FREQ
MVR_PTS
OLDCLAIM
TARGET_AMT
value
90000
60000
30000
0
0
10
20
0
1
2
3
4
5
0
5
10
20000
40000

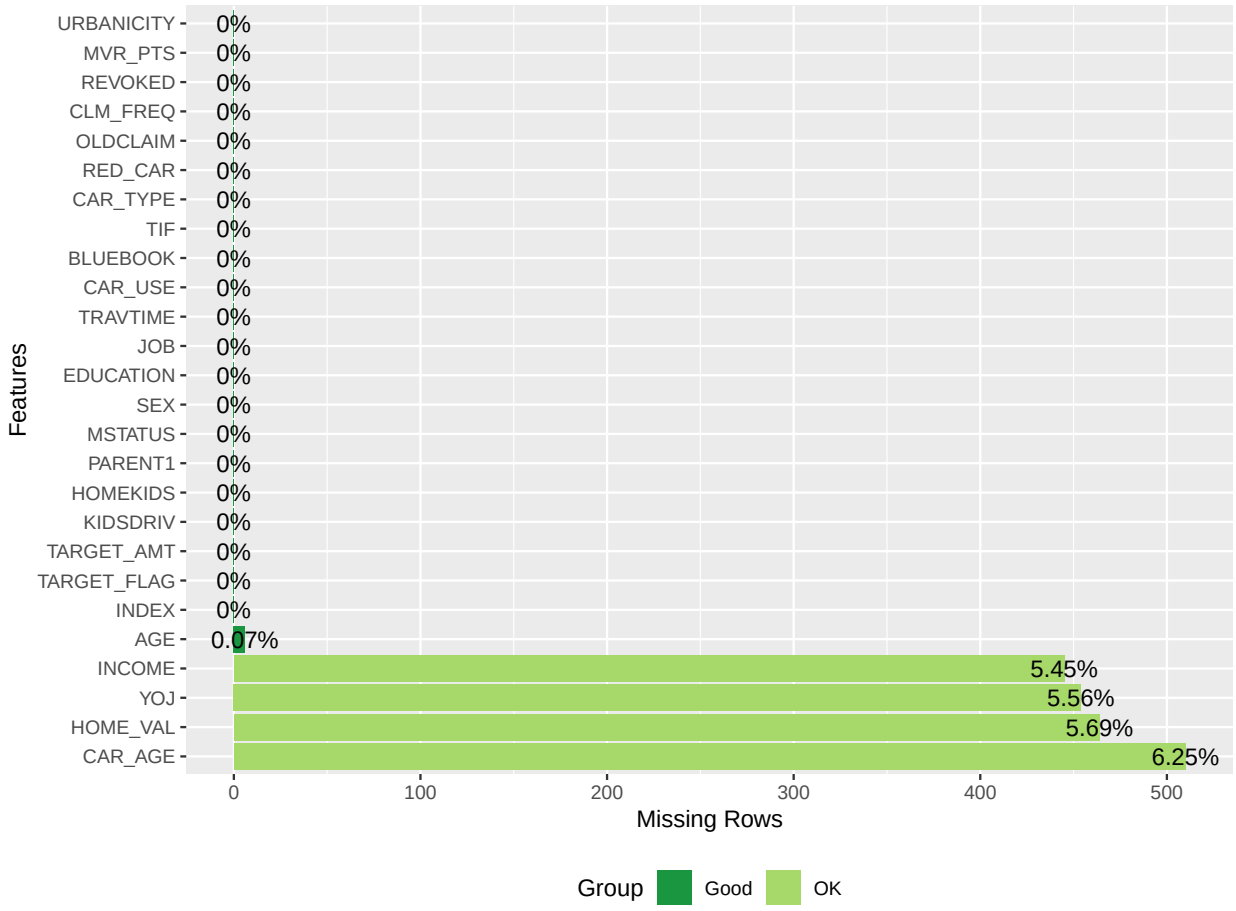


Figure 3: Missing data

There are a few missing data: `AGE`, `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`. Given the low proportion, it seems acceptable to impute the missing values.

2 DATA PREPARATION

2.1 Variable Desc

2.1.0.1 KIDSDRIV

`KIDSDRIV` is a discrete predictor with values ranging from 0 to 4. It shows heavy skewness with most cars having 0 kid drivers. Judging from the distribution, it appears that having kid driver results in higher probability of making a claim.

2.1.0.2 AGE

`AGE` presents driver's age and shows normal distribution, centered around 45. Looking at the boxplot of age, there is no difference between the claim made or not in distribution. Therefore, we can believe that `AGE` may not be helpful in determining the probability of making a claim.

2.1.0.3 HOMEKIDS

`HOMEKIDS` is a predictor describing number of children at home ranging from 0 to 5.

2.1.0.4 YOJ

`YOJ` is a predictor describing years on job. It is believed that people who stay at a job for a long time are usually more safe. YOJ shows normal distribution apart from those who are unemployed.

2.1.0.5 INCOME

`INCOME` is a heavily skewed predictor variable. The outliers should be treated.

2.1.0.6 HOME_VAL

`HOME_VAL` is a home value predictor variable. In theory, home owners tend to drive more responsibly. In the graph, we can see difference between the owners and renters.

2.1.0.7 TRAVTIME

`TRAVTIME` is a predictor variable describing the distance to work. Long drives to work usually suggest greater risk. However the graph shows fairly normal distribution and it may not be helpful determining the probability of making a claim.

2.1.0.8 BLUEBOOK

`BLUEBOOK` is a predictor variable describing the value of the car. The boxplot shows that the lower value of the car, the higher chances of making a claim. It is a possibility that the higher price cars are driven more carefully.

2.1.0.9 TIF

`TIF` describes how long the customer has been with the company, and the longer they have, the safer it may be. The plots show the safe drivers tend to stay safe.

2.1.0.10 OLDCLAIM

`OLDCLAIM` is a predictor describing the claims cost made in the past 5 years. We can see that it is very heavily skewed and that most people do not make claims.

2.1.0.11 CLM_FREQ

`CLM_FREQ` is a predictor that describes claim costs in the past 5 years. It seems that people who have made a claim in the past 5 years are highly likely to make another claim.

2.1.0.12 MVR_PTS

`MVR_PTS` is a predictor that describes motor vehicle record points. If you get lots of traffic tickets, you tend to get into more crash. It appears to be a highly significant variable as seen in boxplots.

2.1.0.13 CAR_AGE

`CAR_AGE` describes the vehicle age. There is one data point that shows the vehicle age is -3, this will be corrected to 0.

2.1.0.14 PARENT1

`PARENT1` describes single parent. This is factorized and renamed as `NumParents` to describe the number of parents.

2.1.0.15 SEX

`SEX` describes the gender of the driver. This is factorized and renamed as `MALE` to describe male as 1 and female as 0. It does not appear to be significant variable in the box plot.

2.1.0.16 MSTATUS

`MSTATUS` describes the martial status of the driver. It is believed that married people drive more safely. This variable has been factorized and renamed as `Single` to explain married as 0, not married as 1.

2.1.0.17 EDUCATION

`EDUCATION` describes the education level of the driver. It is factorized. It may be correlated with INCOME.

2.1.0.18 JOB

`JOB` describes the type of job the driver has. It is factorized. It may be correlated with INCOME. In theory white collar jobs tend to drive safer.

2.1.0.19 CAR_TYPE

`CAR_TYPE` describes type of car. It is factorized.

2.1.0.20 CAR_USE

`CAR_USE` describes how the car is used. Commercial vehicles are driven more and may increase probability of collision. It is factorized and renamed as `Commercial`. 0 means private.

2.1.0.21 RED_CAR

`RED_CAR` describes the color of the car is red. It is believed that red cars, especially sports cars are riskier. It is factorized.

2.1.0.22 REVOKED

`REVOKED` describes whether the license has revoked in the past 7 years. If it has revoked, it shows you are a risky driver. It is factorized. The boxplot shows the drivers who had lost their license are likely to be in accidents.

2.1.0.23 URBANICITY

`URBANICITY` describes whether driver lives in Urban area or Rural area. It is factorized and renamed as `URBAN`. 0 means rural.

2.2 Missing values

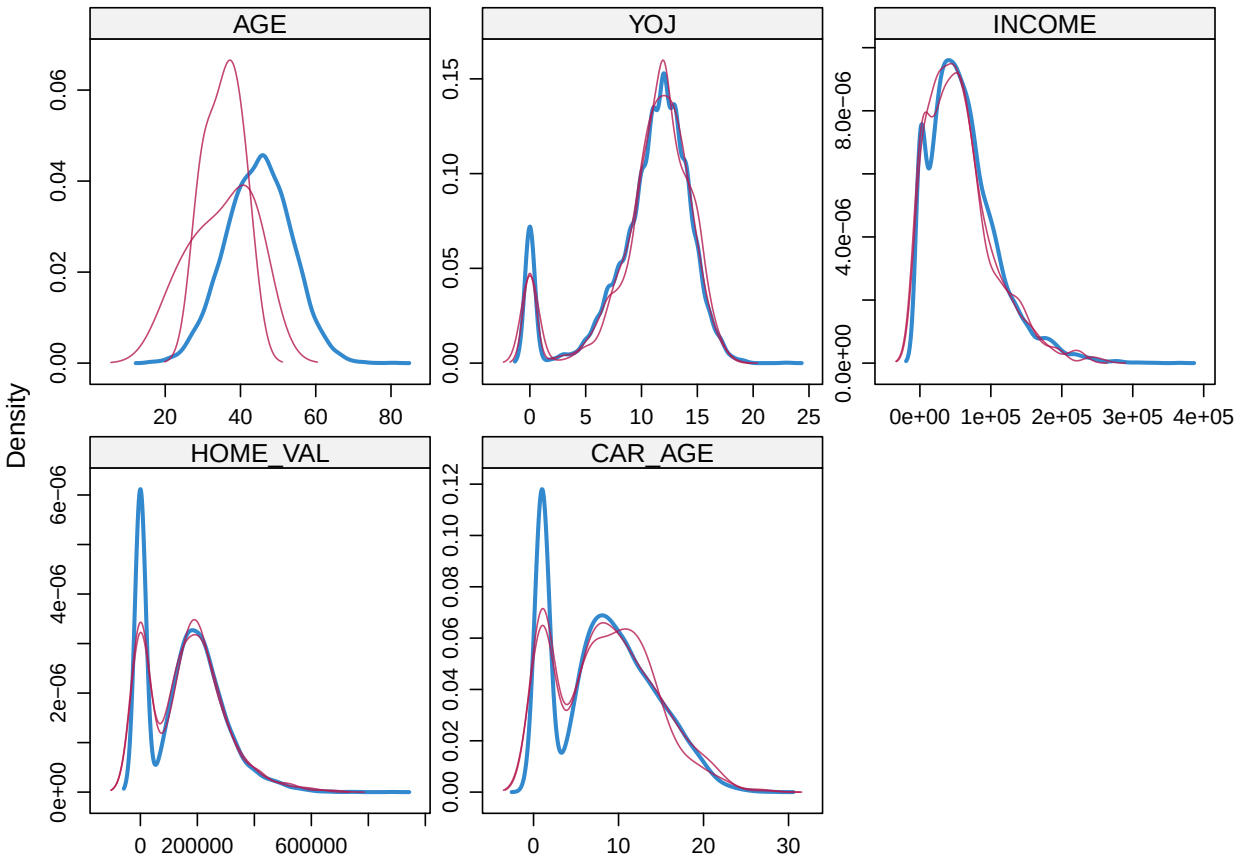
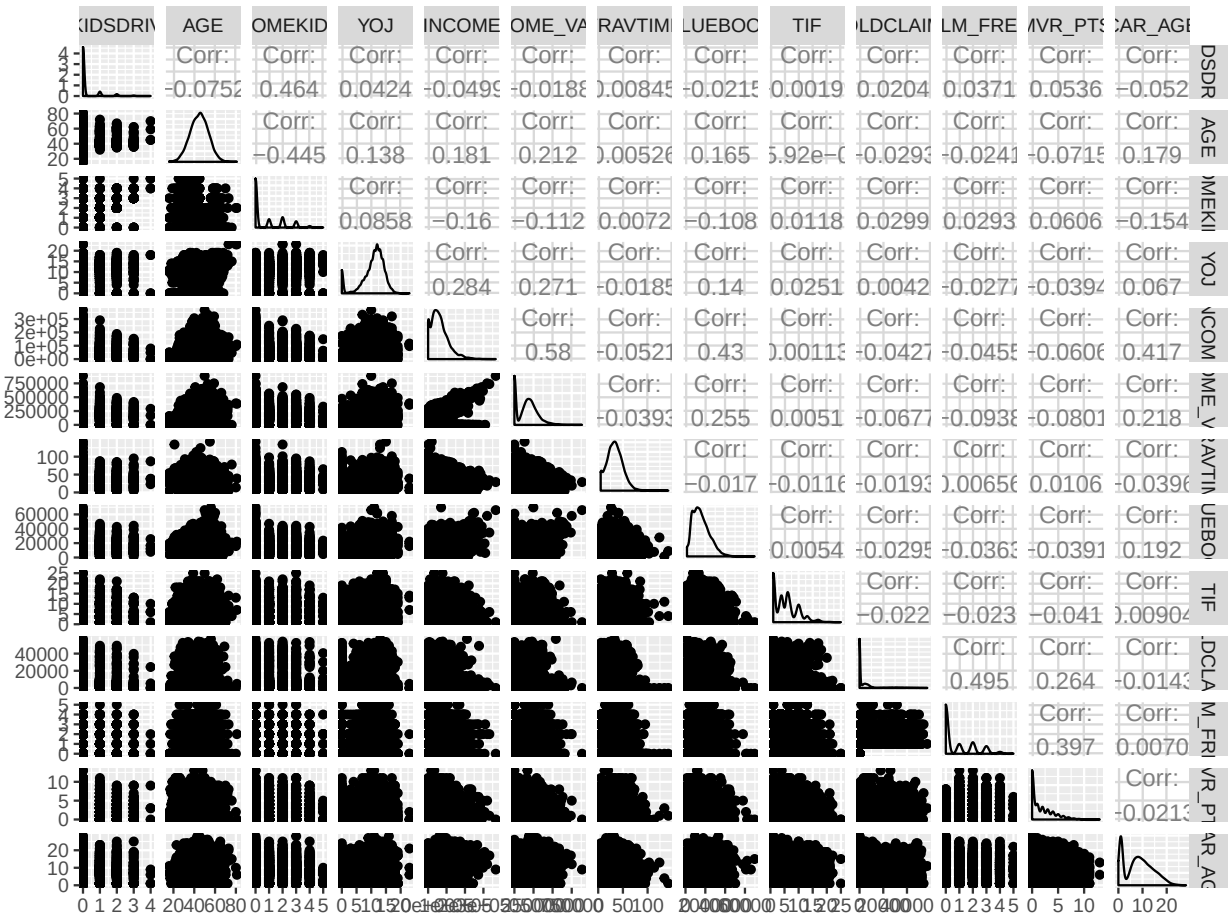
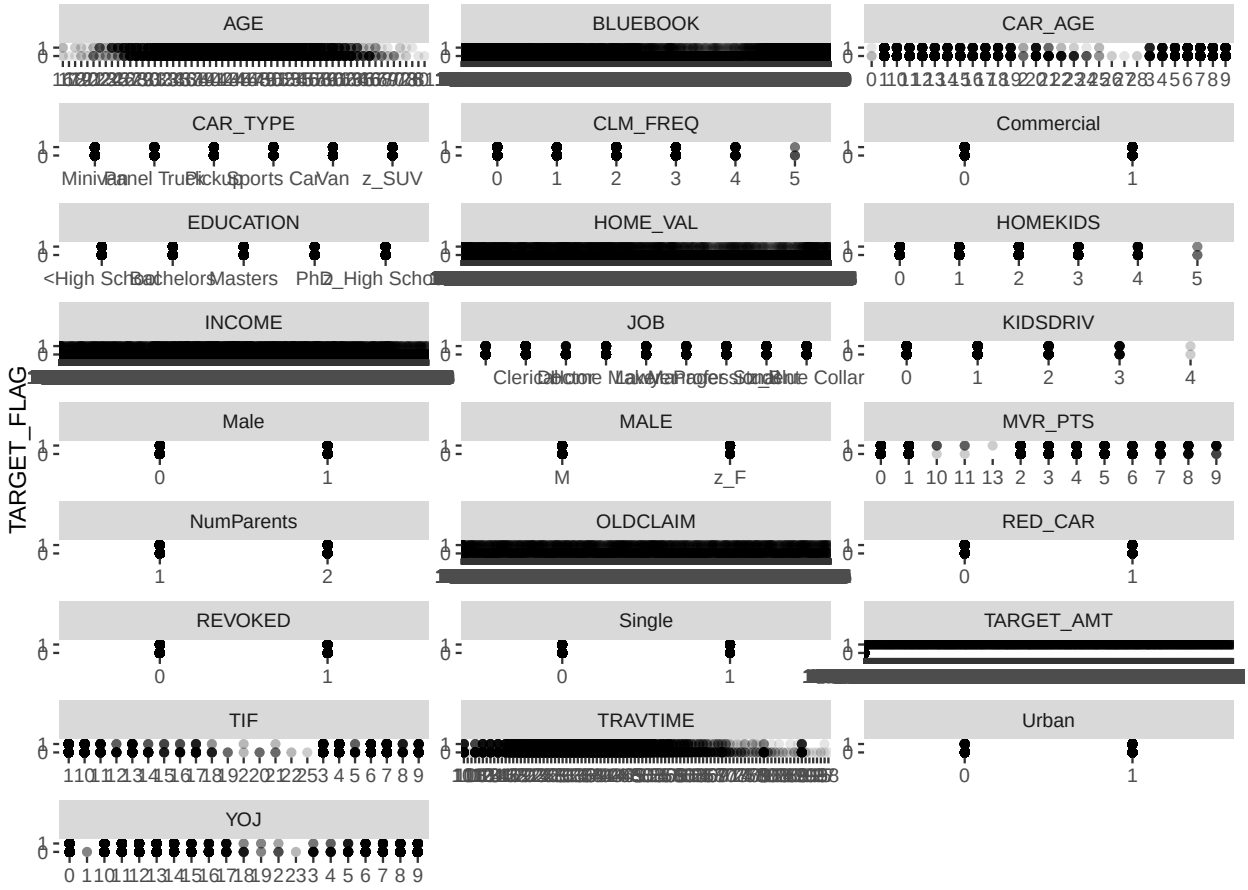


Figure 4: Density plot of imputed data

We can see that excpet the `AGE`, the 4 variables roughly matches the existing distribution. We will use the 4 variables and impute `AGE` separately, using the median imputation.


KIDSDRIV
AGE
OMEKID
YOJ
INCOME
OME_VA
RAVTIM
LUEBOC
TIF
LDCLAIM
LM_FRE
MVR_PTS
CAR_AGE
Corr:
−0.0752
0.464
0.0424
−0.0499
−0.0188
0.00845
−0.0215
0.0019
0.0204
0.0371
0.0536
−0.052
−0.445
0.138
0.181
0.212
0.00526
0.165
−0.0293
−0.0241
−0.0715
0.179
0.0858
−0.16
−0.112
0.0072
−0.108
0.0118
0.0299
0.0293
0.0606
−0.154
0.284
0.271
−0.0185
0.14
0.0251
0.0042
−0.0277
−0.0394
0.067
0.58
−0.0521
0.43
0.00113
−0.0427
−0.0455
−0.0606
0.417
−0.0393
0.255
0.0051
−0.0677
−0.0938
−0.0801
0.218
−0.017
−0.0116
−0.0193
0.00656
0.0106
−0.0396
0.0054
−0.0295
−0.0363
−0.0391
0.192
−0.022
−0.023
−0.041
0.00904
0.495
0.264
−0.0143
0.397
0.0070
−0.0213




```
##           KIDSDRIV   AGE HOMEKIDS   YOJ INCOME HOME_VAL TRAVTIME BLUEBOOK
## KIDSDRIV      1.00 -0.08     0.46  0.04  -0.05    -0.02      0.01    -0.02
## AGE          -0.08  1.00    -0.45  0.14   0.18     0.21      0.01     0.16
## HOMEKIDS      0.46 -0.45     1.00  0.09  -0.16    -0.11     -0.01    -0.11
## YOJ           0.04  0.14     0.09  1.00   0.28     0.27     -0.02     0.14
## INCOME       -0.05  0.18    -0.16  0.28   1.00     0.58     -0.05     0.43
## HOME_VAL     -0.02  0.21    -0.11  0.27   0.58     1.00     -0.04     0.26
## TRAVTIME      0.01  0.01    -0.01 -0.02  -0.05    -0.04      1.00    -0.02
## BLUEBOOK     -0.02  0.16    -0.11  0.14   0.43     0.26     -0.02     1.00
## TIF           0.00  0.00     0.01  0.03   0.00     0.01     -0.01    -0.01
## OLDCLAIM      0.02 -0.03     0.03  0.00  -0.04    -0.07     -0.02    -0.03
## CLM_FREQ      0.04 -0.02     0.03 -0.03  -0.05    -0.09      0.01    -0.04
## MVR_PTS       0.05 -0.07     0.06 -0.04  -0.06    -0.08      0.01    -0.04
## CAR_AGE      -0.05  0.18    -0.15  0.07   0.42     0.22     -0.04     0.19
##                TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE
## KIDSDRIV  0.00      0.02     0.04     0.05   -0.05
## AGE       0.00     -0.03    -0.02    -0.07    0.18
## HOMEKIDS  0.01      0.03     0.03     0.06   -0.15
## YOJ       0.03      0.00    -0.03    -0.04    0.07
## INCOME    0.00     -0.04    -0.05    -0.06    0.42
## HOME_VAL  0.01     -0.07    -0.09    -0.08    0.22
## TRAVTIME -0.01     -0.02     0.01     0.01   -0.04
## BLUEBOOK -0.01     -0.03    -0.04    -0.04    0.19
## TIF       1.00     -0.02    -0.02    -0.04    0.01
```

```
## OLDCLAIM -0.02     1.00     0.50    0.26   -0.01
## CLM_FREQ -0.02     0.50     1.00    0.40   -0.01
## MVR_PTS  -0.04     0.26     0.40    1.00   -0.02
## CAR_AGE   0.01    -0.01    -0.01   -0.02    1.00
```

3 BUILD MODELS

4 SELECT MODELS

5 Appendix

The appendix is available as script.R file in `project4_insurance` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining